Knowledge Discovery & Dissemination

Enabling Analysts To Quickly Produce Actionable Intelligence From Multiple Sources Of Information

Dr. Arthur H. Becker

KDD Proposers Day

IARPA-BAA-09-10

20 August 2009

Purpose

To Present Current Ideas And Thoughts On The KDD Program

To Solicit Comments And Questions From Potential Bidders

To Provide A Forum For Teaming And Collaboration

Disclaimer

- This presentation is provided solely for information and planning purposes
- The Proposers' Day Conference does not constitute a formal solicitation for proposals or proposal abstracts
- Nothing said at Proposers' Day changes the requirements set forth in a BAA
- Any conflict between what is said at Proposers' Day and what is in a BAA will be resolved in favor of the BAA

Overview Topics

KDD Research

KDD Program Overview – The 12 Month Cycle

Programmatic Topics

Note: Questions During Proposers' Day :

- Answers will be provided verbally by the briefer
- Questions and answers will be recorded and posted
- The posted answers will be the official response
- Additional questions can be submitted through 27 August

KDD Research

What Are We Trying To Do?

Objective: Enable Analysts To Quickly Produce Actionable Intelligence From Multiple Sources Of Information

To include newly available data sets that are unfamiliar to the analyst

KDD Research

Automated alignment of data models

Advanced analytic algorithms that work across multiple data sources

KDD Evaluation

Test research against real data and realistic IC problems

Develop challenge problems that drive research toward IC solutions

Research – Alignment of Data Models

Each data set has a data model associated with it. The model could be explicit or implicit, known or unknown

The data model describes how the data is organized and labeled. It may also describe terminology and how elements of the data are related to one another

To make effective use of multiple data sets, the data models need to be aligned so that terms can be used consistently

KDD is interested in innovative research that pushes the envelope in automated data model alignment

Techniques in Ontology Alignment, Machine Learning, Language Processing and Folksonomic data models are among the research areas applicable to this problem

Alignment Example

Consider a data base of incident reports. Terminology used in the data base may be different than those of the analyst

Data Base Terminology	Analyst Terminology
Location (Address)	Location (Lat/Long)
Locality (Province)	n/a
Region (Area)	Region (Province)
Incident Type (Types of Crimes)	Incident Type (Types of Attacks)
-- Arson	-- Kinetic
-- Hoax	-- Biological
-- Murder	-- Radiological
-- Bombing	-- Chemical
	-- Cyber

Research - Advanced Analytic Algorithms

Human Cognition **(Requires Interpretation and Strategy)**	• What is X's capability to produce BW
	• What are the intentions of Country Y
	• Is this movement gaining strength
Advanced Analytics **(Requires a Data Model)**	• What are the topics of these documents
	• Find documents similar to this one
	• Find more documents about Topic X
	• Identify groups in this network
	• Who are the key members, their roles
	• Find people like this
	• What is near this location
	• What events, people are near this location
	• What's changed recently
Direct Queries **(No Model Required)**	• Find documents with the string "XYZ"
	• Who did Person X contact
	• Pull the record on John Doe

Research – Advanced Analytic Algorithms Across Multiple Data Sets

Advanced analytic algorithms can perform more complex analytic tasks by understanding (to some extent), the analyst's interests

The algorithms understand the meaning of relevance, significance, similar, important …etc

Advanced analytic algorithms have been developed for single data sets or multiple data sets where the data is very similar

KDD is interested in research to develop advanced analytic algorithms that work across data sets that are very different

Types of advanced analytic algorithms include, but are not limited to:

- Semantic Distance (used for topic clustering, topic spotting and example based search)
- Social Network Analysis using content and multiple types or relationships
- Entity resolution and merger of geospatial and network data

Research Not Of Interest

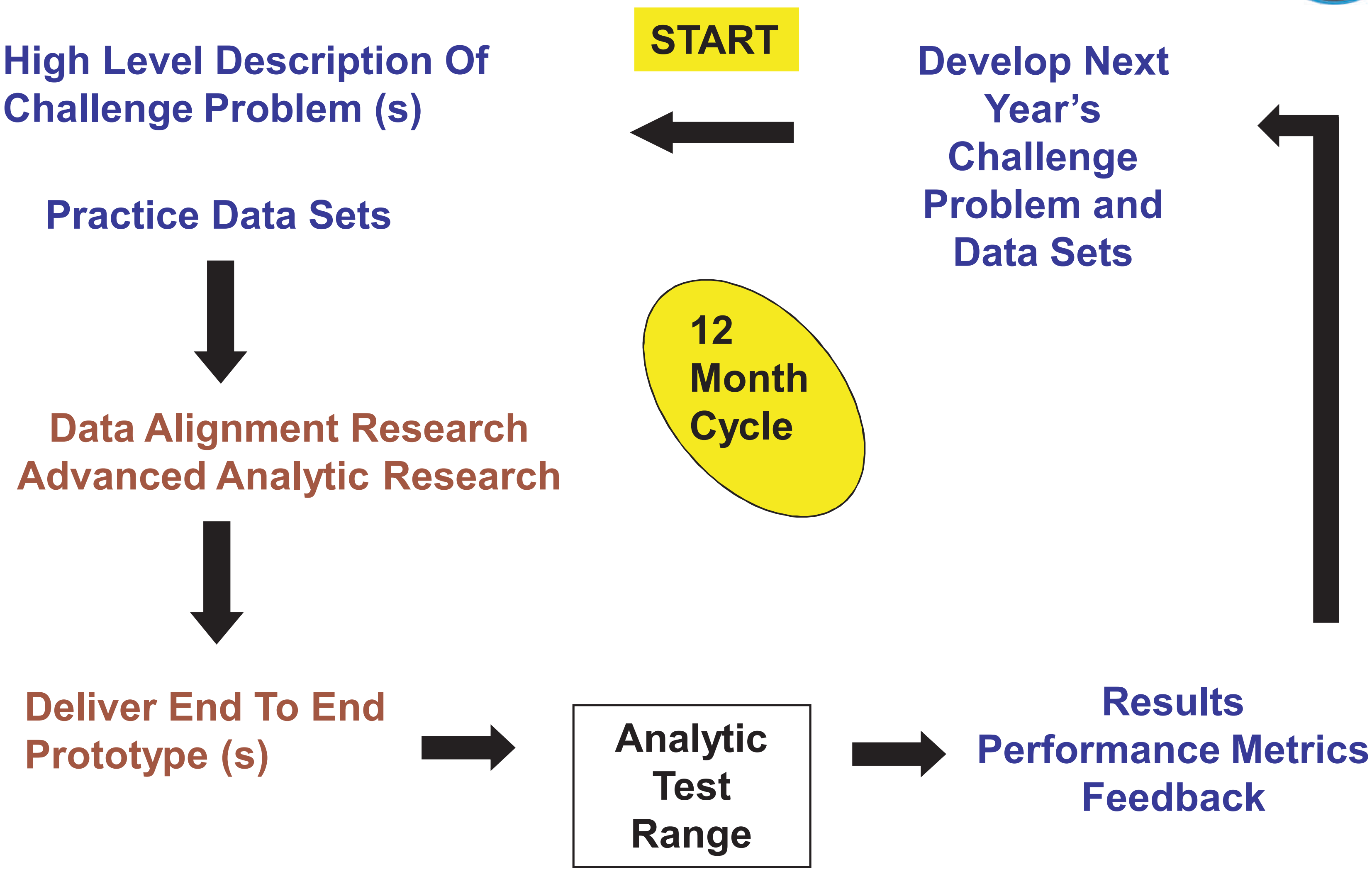
- Studies of analyst behavior or analytic teaming
- Processing media formats (Speech, Video …etc)
- Hypothesis generation, hypothesis validation and sense making
- Alternate reporting methods such as storytelling and video generation
- Machine translation and foreign language processing
- Research in visualization technology
- Specialized hardware for analytic processing
- Computer architecture research for analytic processing
- Natural language interfaces between the analyst and analytic tools

Questions?

KDD Program Overview

The 12-Month Cycle

KDD Program Overview



START

High Level Description Of Challenge Problem (s)

Practice Data Sets

Data Alignment Research
Advanced Analytic Research

Deliver End To End Prototype (s)

Analytic Test Range

Results
Performance Metrics
Feedback

Develop Next Year's Challenge Problem and Data Sets

12 Month Cycle

Hypothetical Problem

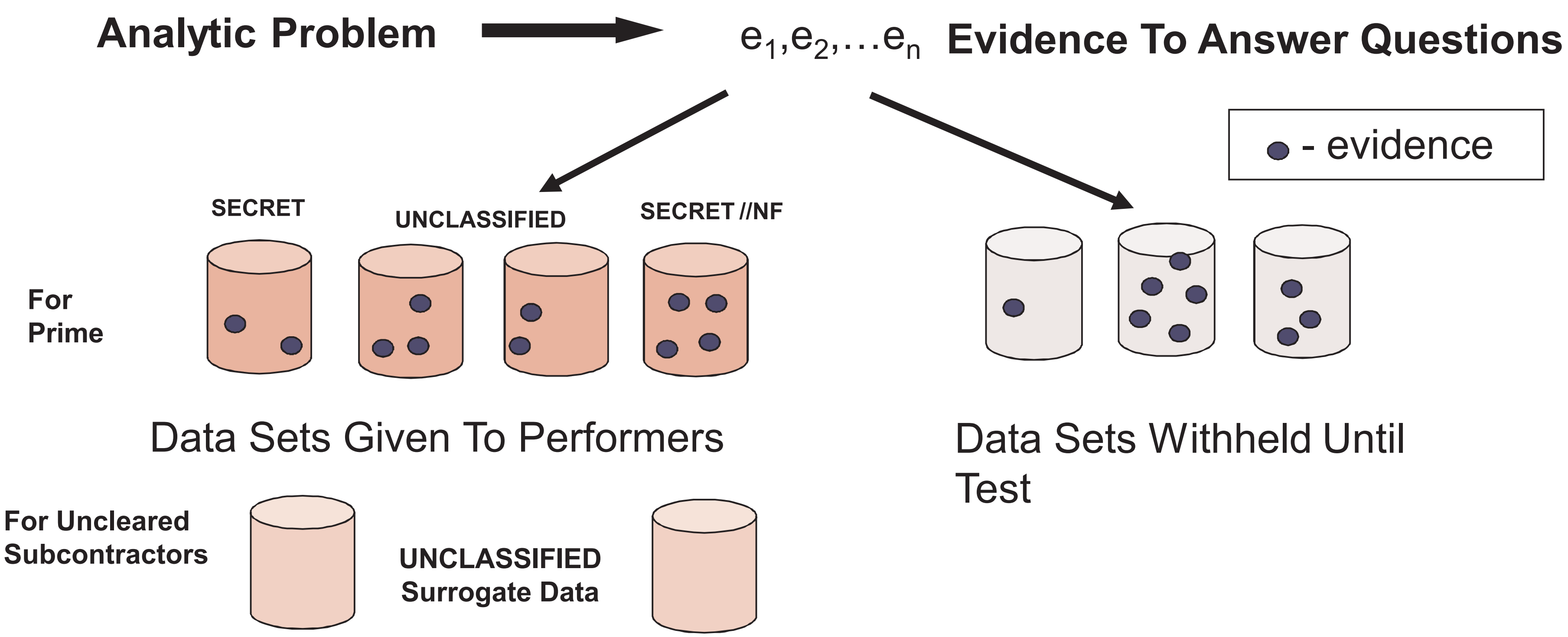
High Level Problem: "Attached is a report indicating the existence of a previously unknown foreign terrorist group. Information about two persons mentioned in the group are attached." Answer the following questions:

The high level problem creates specific advanced analysis tasks:

- **How large is this group?**
- **Who are the key players and what are their roles?**
- **Are there connections to other terrorist organizations?**
- **Are there indications that they have already committed terrorist acts?**
- **Are there specific skills or interests within the group?**
- **What is the primary nationality of the membership ?**
- **Are there other organizations distributing a similar message?**

Direct questions can be answered with fully automated processing to create test probes

Data Sets to Support Analytic Problem



Analytic Problem → $e_1,e_2,\ldots e_n$ **Evidence To Answer Questions**

● - evidence

For Prime

SECRET

UNCLASSIFIED

SECRET //NF

Data Sets Given To Performers

Data Sets Withheld Until Test

For Uncleared Subcontractors

UNCLASSIFIED Surrogate Data

Multiple Problems May Be Used With A Given Suite Of Data Sets

Data Sets Provided to Performers by KDD PMO

Data sets can be structured and unstructured. They can include data bases of different types, flat files, spreadsheets, and more

Data models can be expressed explicitly as a formal ontology or implicitly. We frequently do not know the model or the ontology

Data sets can have a high percentage of missing data, errors and duplications. Data sets could be developed at different times

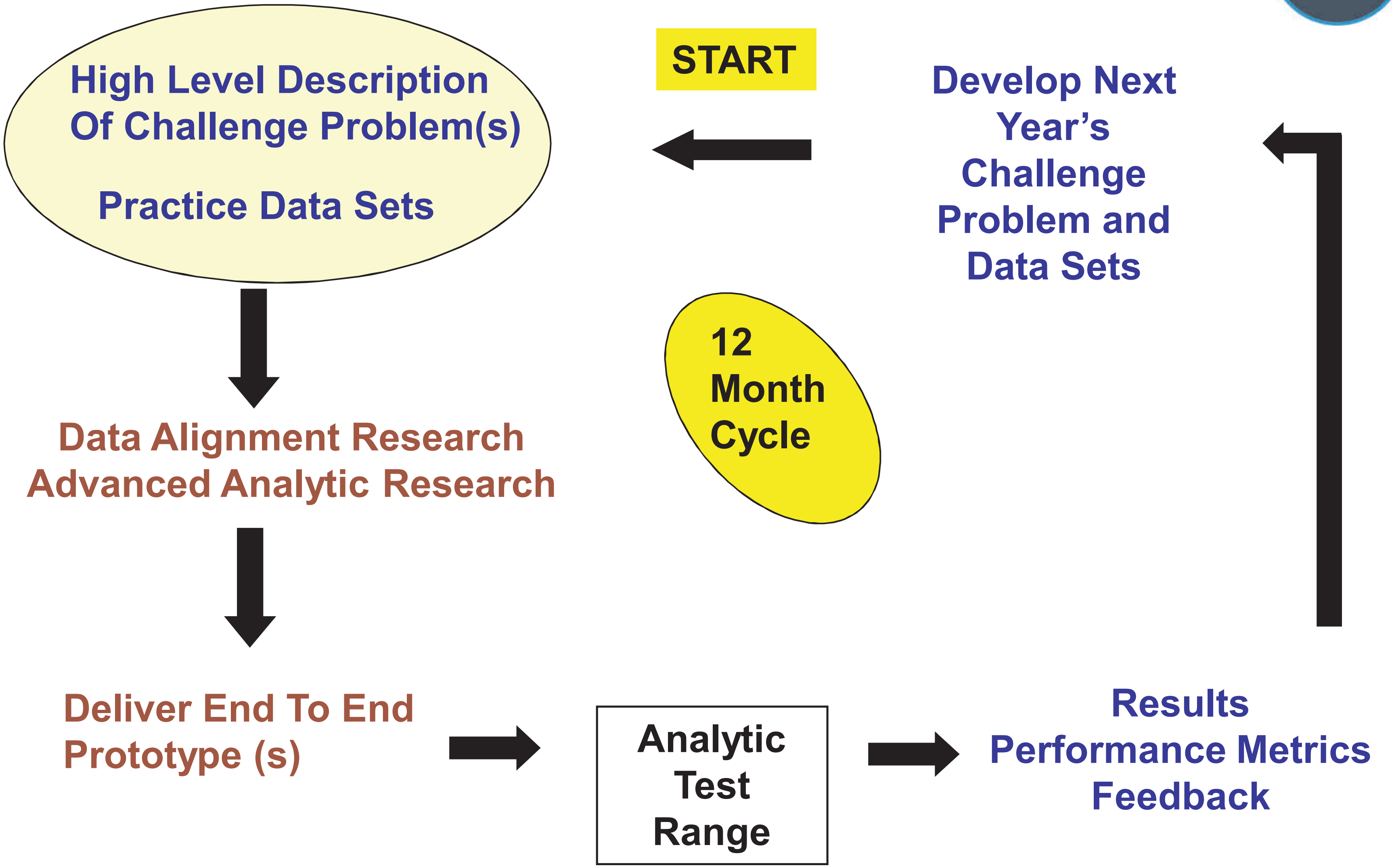
Foreign language text will only be included as addresses, names, and locations. Column headers and spreadsheet items may contain foreign words

The data sets will not require processing speech, video or images. The data sets may be a product of transcription or translation

Data will be real and some data sets will be classified. Classification will be no higher than SECRET// NOFORN. KDD will provide unclassified surrogate data sets to support research by uncleared researchers

Participants are required to conduct all experiments in compliance with applicable laws and policies, including those relating to the use of approved data sets and the protection of the privacy and civil liberties of U.S. Persons

KDD Program Overview



START

High Level Description
Of Challenge Problem(s)

Practice Data Sets

Data Alignment Research
Advanced Analytic Research

Deliver End To End
Prototype (s)

Analytic
Test
Range

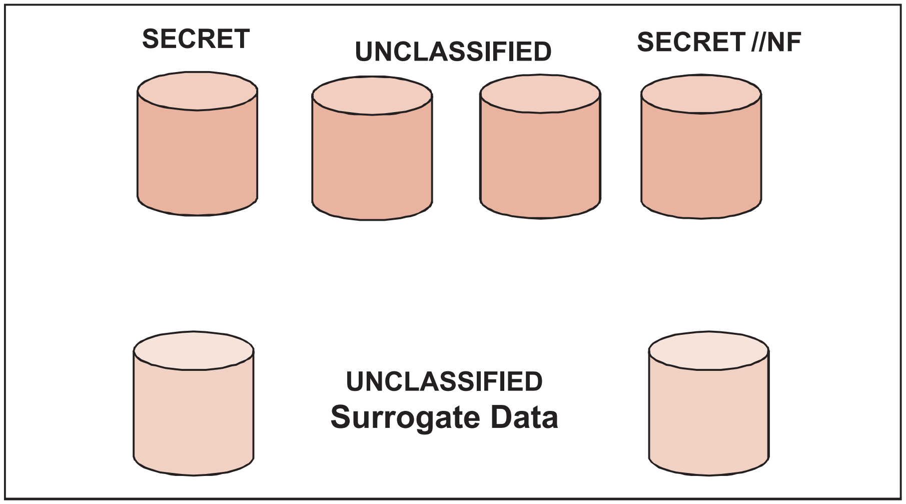
Results
Performance Metrics
Feedback

Develop Next
Year's
Challenge
Problem and
Data Sets

12
Month
Cycle

Provided To Performer At The Beginning Of The Cycle

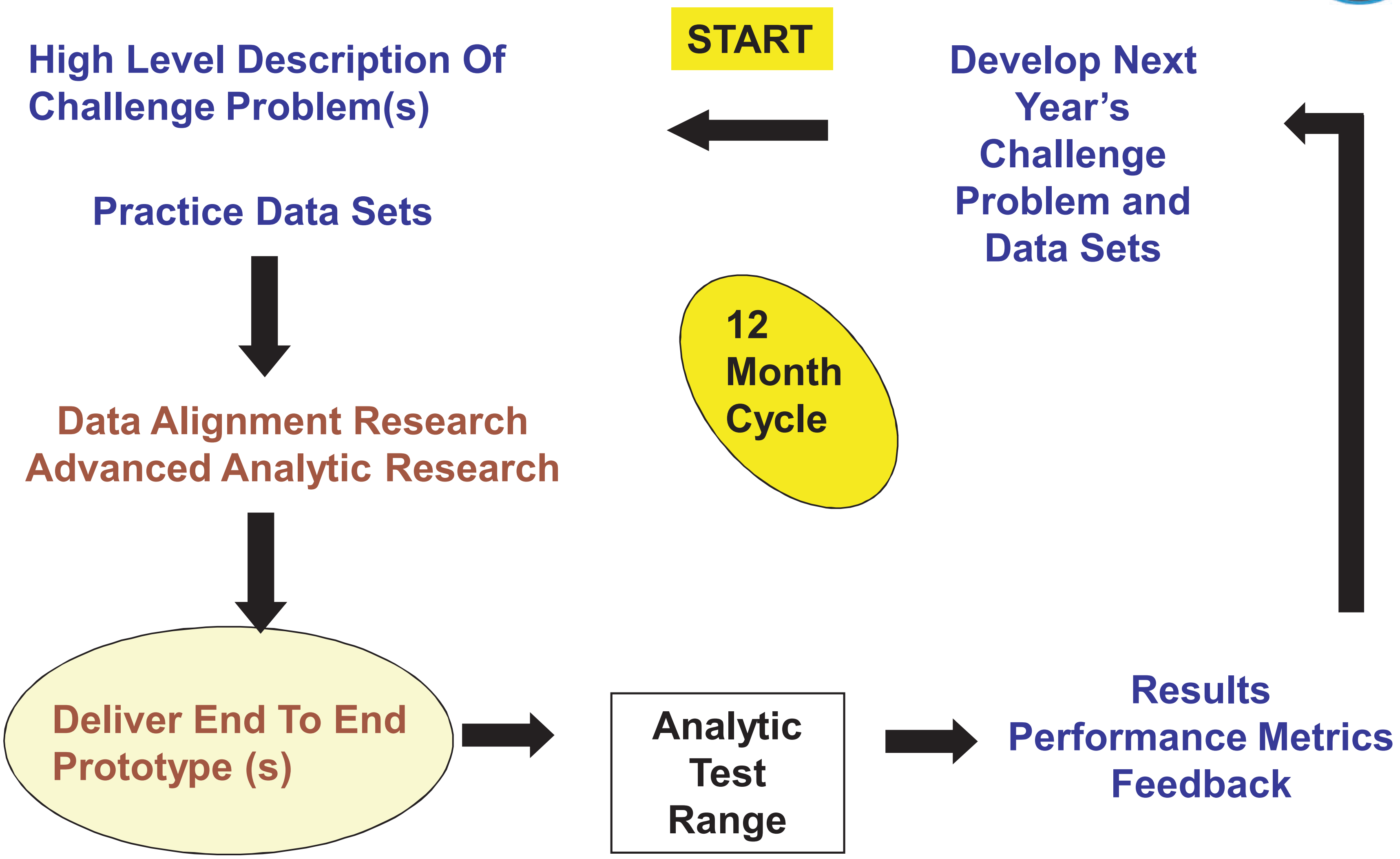
High Level Problem Description: (Based on a hypothetical problem) The problem involves analysis of a foreign group. Some people in the group are known. Analysis will involve determining other members and characterizing the members and activities of the group.



Additional Data Provided At Evaluation: Additional data sets will include 1) A regional data set of biographic data; 2) An internal data set containing Incident Reports; and 3) A data set of translated newspaper articles

Each performer will have up to 200 hours of support from the BLACKBOOK software development team during the first year of the program

KDD Program Overview



START

High Level Description Of Challenge Problem(s)

Practice Data Sets

Data Alignment Research
Advanced Analytic Research

Deliver End To End Prototype (s)

Analytic Test Range

Results
Performance Metrics
Feedback

Develop Next Year's Challenge Problem and Data Sets

12 Month Cycle

Prototype Delivery

- The performer will deliver two prototypes: One for alignment and one for advanced analytic algorithms
- Prototypes will be delivered as source code as well as object code. The T&E team will install into a test suite
- Advanced analytic prototype software will be delivered as services integrated into BLACKBOOK and pre-tested on hardware that will be provided as GFE
- To be determined: Whether alignment prototype software must also run on the GFE hardware, or whether the performer can use other hardware (with appropriate security documentation)
- Output of the alignment prototype must be OWL-2 statements and cannot alter the original data sources or create merged data sources (a data warehouse)
- Performer may elect to bring other supporting data to the evaluation to support alignment
- The alignment prototype may accept feedback from the advanced analytic prototype during the evaluation to support adaptive methods and machine learning

Proposed GFE Hardware

Supermicro 1U Twin SuperServer

1x 6015TW-TB: Supermicro 1U Twin Rackmount SuperServer 6015TW-TB

- Two systems (nodes) in a 1U Form Factor with each node supporting the following:
- Dual Xeon 5400 Series Sockets, 1600Mhz FSB
- Dual Onboard Gigabit Ethernet and Onboard Video
- 1 (x16) PCI-Express Generation II
- Sharing a 980W High-efficiency Power Supply

4x E5420: Harpertown E5420 2.5G 12M 1333FSB

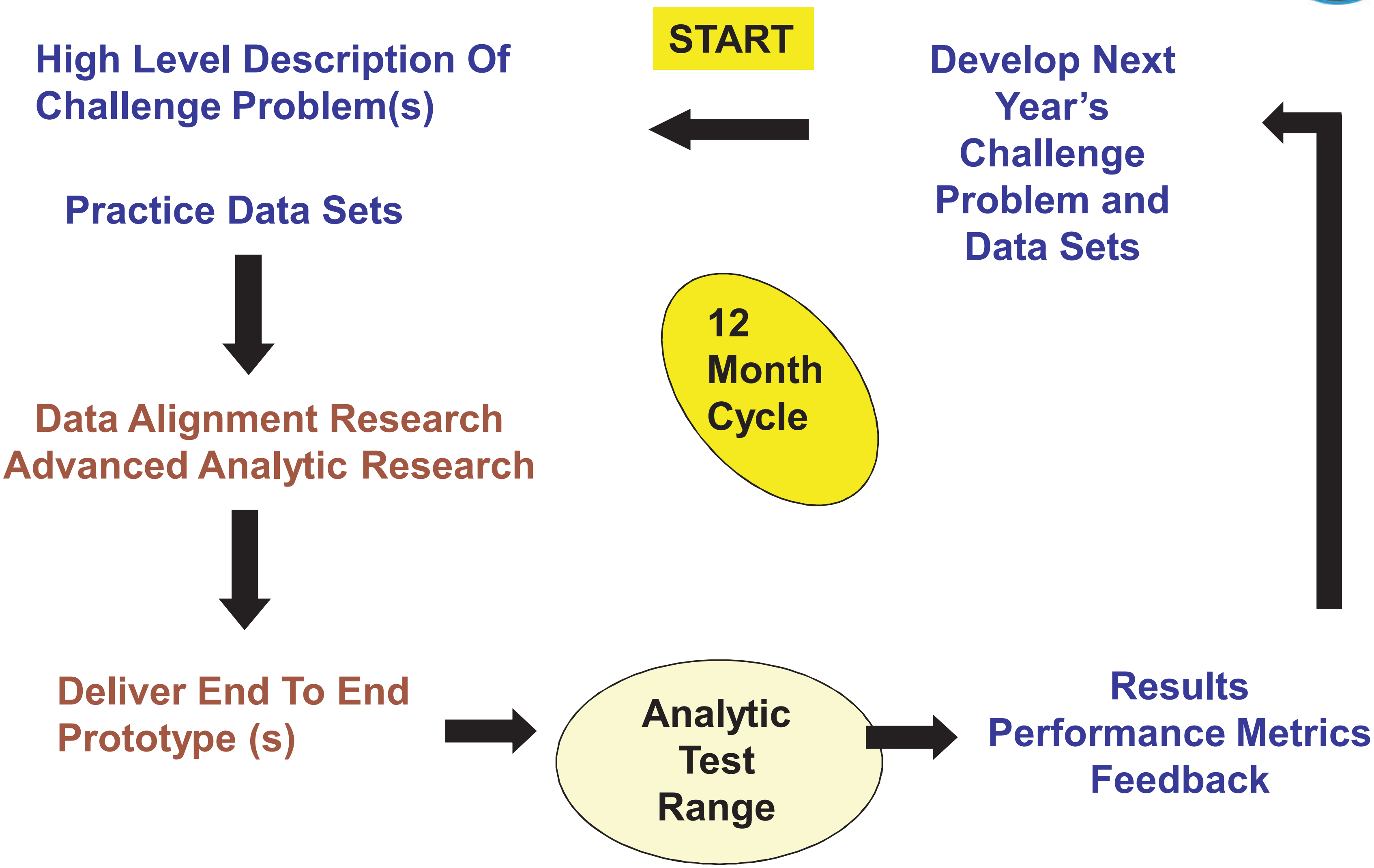
8x ACT2GFR72M8G667M4: 4GB 667Mhz ECC FBD (ACT4GFR72M8G667S)

4x ST31000340NS: SEAGATE BARRACUDA ES.2 - HARD DRIVE - 1TB INTERNAL - 3.5IN - SATA 3GB/S - 7200R

2x AOC-SIMSO+: IPMI 2.0 with Virtual Media Over LAN & Dedicated LAN

Mini USB with 8" Cable for Dedicated LAN

KDD Program Overview



START

High Level Description Of Challenge Problem(s)

Practice Data Sets

Data Alignment Research
Advanced Analytic Research

Deliver End To End Prototype (s)

Analytic Test Range

Results
Performance Metrics
Feedback

Develop Next Year's Challenge Problem and Data Sets

12 Month Cycle

Evaluation: The Analytic Test Range

The centerpiece of KDD's evaluation plan is an analytic test range that will measure the performers' prototypes against realistic analytic tasks, involving real data and using IC analysts

When the evaluation begins, performers will be provided data sets and a number of specific analytic tasks to perform

- Some of the data sets will be familiar. They are similar to the data provided at the beginning of the cycle.
- Some of the data sets will be new
- Evidence to answer the analytic tasks will be seeded in the data sets and will serve as ground truth

Performers will have a fixed amount of time to align the data sets using their alignment prototype, develop automatic probes for specific tasks and prepare their advanced analysis prototype

- Automatic probes are for analytic tasks that can be performed without analyst intervention using a workflow process

Analysts will then use the performer's advanced analysis prototype to perform the remaining analytic tasks (not automatic probes)

Evaluation: (continued)

At fixed times, analysts' results will be collected and recorded.

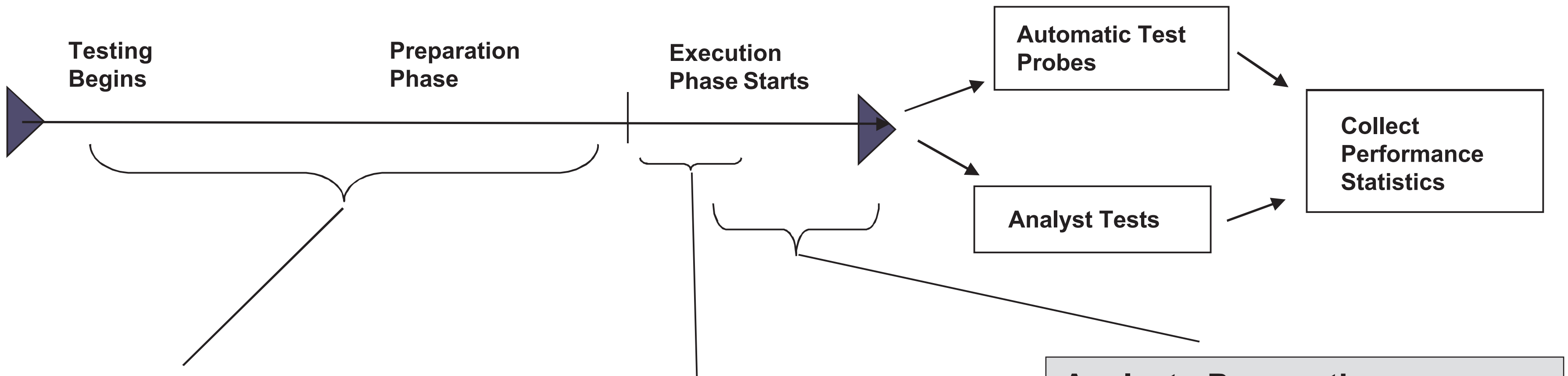
The performers' prototypes will be measured on how quickly, completely and accurately the analysts used their prototype to perform the analytic tasks

Completeness and accuracy will use standard precision and recall measures against ground truth

Note: The performers' algorithms will not be measured separately (e.g. how well their alignment algorithm performed); but rather, on how well each performer's prototype system performed the overall end-to-end task

At each evaluation, several test problems may be used over the same data sets. Analysts will rotate to a different performer's prototype for each problem

Running the Prototypes on the Analytic Test Range



Performers Given:
Surprise data sets

Detailed descriptions of the problem and analytic tasks

Performers Will:
Have limited amount of time to run their alignment prototype on all of the data sets

Performers Given:
Specific probe Questions

Performers Will:
Develop their automatic program to answer probe questions using BLACKBOOK workflow and their aligned data sets

Analysts Preparation:
Receiving training on performers prototype systems

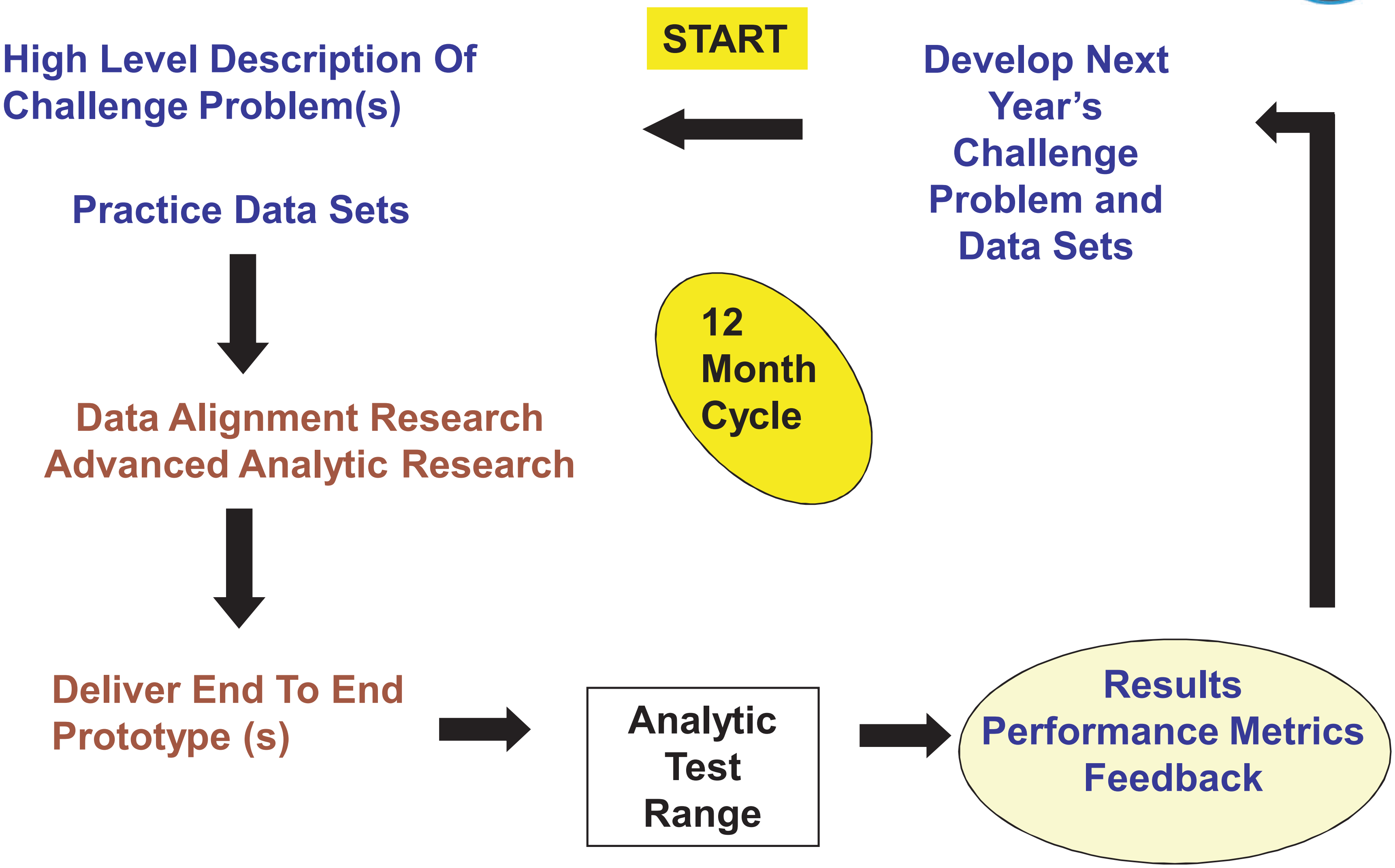
Briefed on the problem

Assigned a specific system to test

Performers may be present during the test to answer technical questions about their prototype systems

Each Evaluation Will Run Multiple Problems. Analysts Will Rotate To Other Prototypes

KDD Program Overview



START

High Level Description Of Challenge Problem(s)

Practice Data Sets

Data Alignment Research
Advanced Analytic Research

Deliver End To End Prototype (s)

Analytic Test Range

Results
Performance Metrics
Feedback

Develop Next Year's Challenge Problem and Data Sets

12 Month Cycle

Metrics

- Metrics are based on the performance of the prototype over the analytic test range. Prototypes will be measured in terms of how accurately, completely and quickly they perform tasks
- Alignment time will be restricted and reduced in later cycles
- All tests will be objective, repeatable and statistically valid
- Statistical validity will be accomplished by use of sufficient number of analysts
- Automatic probe answers will be scored in terms of recall and precision
- BLACKBOOK will be instrumented to collect detailed performance data
- At 12 months an evaluation will be conducted. This “Pre-Test” will validate the evaluation processes and metrics

Post Evaluation

- Post evaluation discussions will occur with each performer team at the prime's facility to review performance results and discuss actions to be taken for next cycle

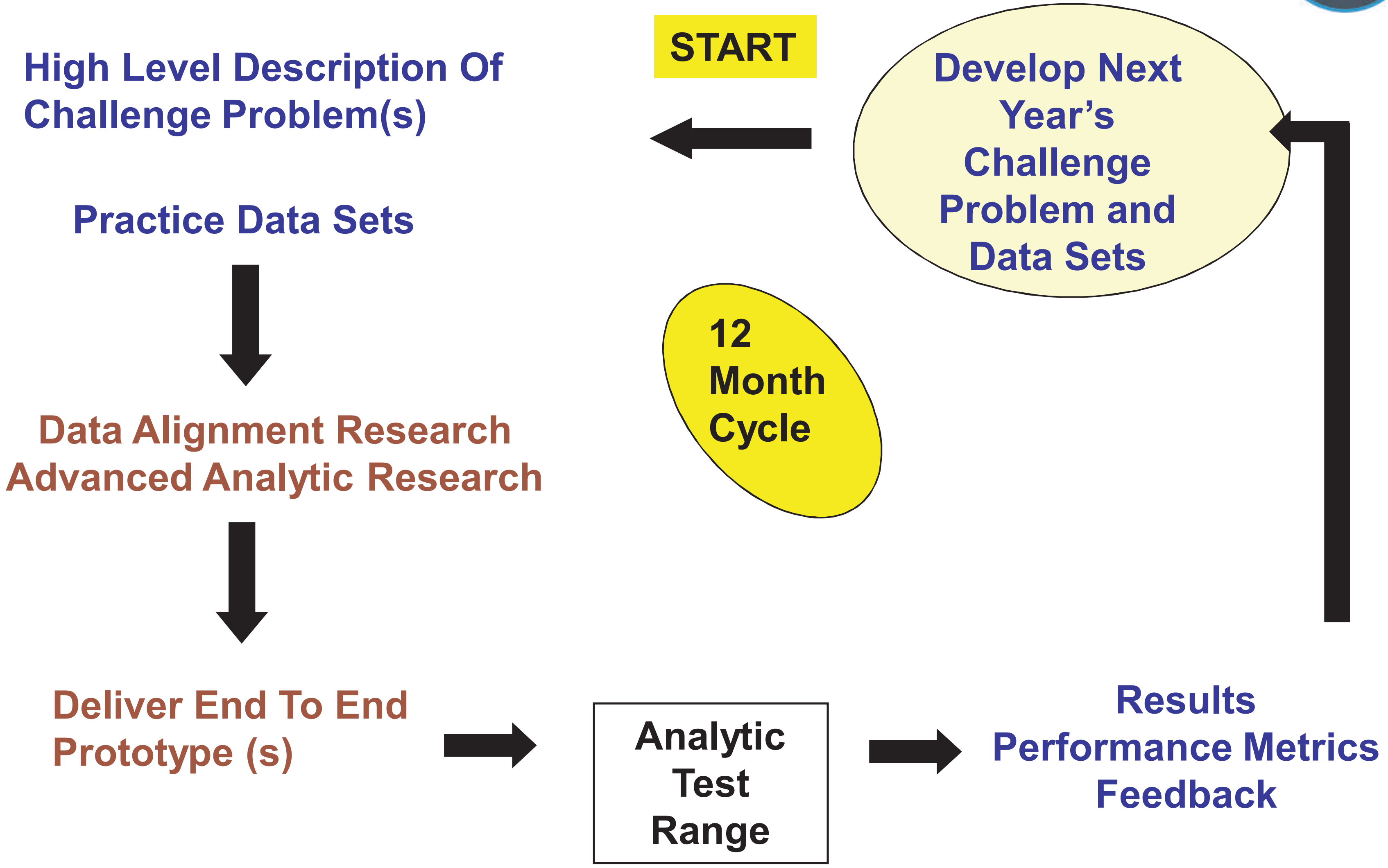
- Performers present their research at the KDD workshop with all performers, Program Office and T&E team in attendance

- Downselection decisions will be based on the quality, progress and aggressiveness of the performer's research and on availability of funds

- Starting at Mid-Term (approximately month 24), downselection will also be based on the performance of the prototype. Performance results during the Pre-Test (at month 12) will not be considered in downselection

- Program Office prepares for next cycle

KDD Program Overview



START

High Level Description Of Challenge Problem(s)

Practice Data Sets

Data Alignment Research
Advanced Analytic Research

Deliver End To End Prototype (s)

Analytic Test Range

Results
Performance Metrics
Feedback

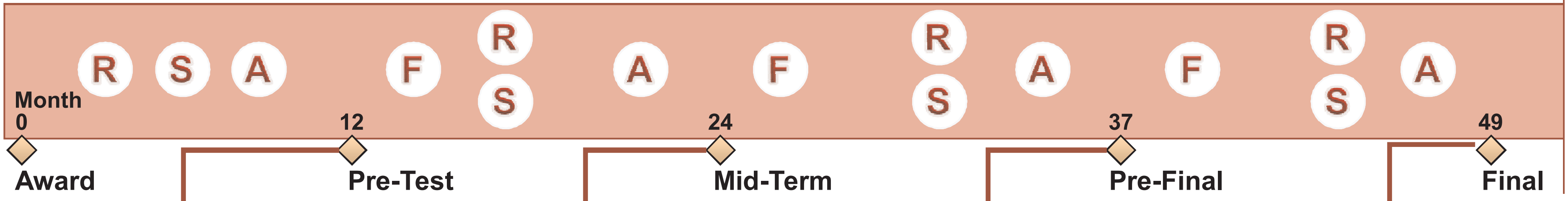
Develop Next Year's Challenge Problem and Data Sets

12 Month Cycle

Develop Next Year's Challenge Problem

- After each test cycle, a new challenge problem will be developed and new data sets will be identified
- Data sets will be provided to performers and unidentified (surprise) data sets will be selected
- Unclassified surrogate data sets will also be developed by the KDD PMO and provided to performers
- Test plans will be refined as necessary and provided

Performer Research Evaluation Timeline



Pre-Test - 12 Months

- Full test with multiple problem sets, but with fewer analysts

Purpose:

- Assess assumptions of test process
- Provide feedback to performers
- Refine parameter estimates for Mid-Term Test
  -- Variance measures
  -- Assess program instrumentation process
  -- Inter-analyst variance
  -- Time required for each problem
  -- Need to address outliers
  -- Alignment time established

Mid-Term - 24 Months

- Full test of each performer's system with enough analysts for statistical significance
- Test each performer's solution against a baseline system
- Test a maximum of two combinations selected by T&E team
- Alignment time restricted

Purpose:

- Assessment of KDD's progress in meeting program goals
- Support program management decisions

Pre-Final - 37 Months

- Full test with multiple problem sets, with enough analysts for statistical significance

Purpose:

- Determine if program goals are being met
- Provide feedback to performers
- Support program management decisions

Final - 49 Months

- Full Test with enough analysts for statistical significance
- Use problems aimed at technology transition partners
- Full test against those systems selected by the T&E team
- Alignment time further restricted

Purpose:

- Determine if KDD has met program goals
- Prepare for technology transfer

Waypoints

R Demonstration of Research in performer's facility

S Demonstration of performer's System in their facility

A Acceptance of performer's system for test

F Feedback and discussion of performers results

Questions?

Programmatic Topics

Teaming

- Teaming is encouraged; however:
  - Teams should have clear, strong management and single point of contact
  - Each team member should contribute significantly to the program goals
  - It should not consist of loose confederations
  - There should be no teaming for teaming sake
- "Ideal Team"
  - A prime contractor that has technical expertise and participates in the research as well as the prototype development and management
  - Subcontracts and partnerships with academic institutions and companies with expertise to complement the prime
  - Researchers with known expertise in the areas KDD is pursuing
- Prime must have personnel cleared at S//NF level and a facility clearance to process and store S//NF material at the time of proposal submission

Award Information

- Four-year program starting in Q1 FY 2010 (Dec 2009)
  - Base Contract of 15 months
  - Option periods of 12 months each

- Multiple awards are anticipated

Eligibility Information

- Collaborative efforts/teaming is strongly encouraged

- Foreign participants and/or individuals may participate
  - Must comply with Non-Disclosure Agreements, Security Regulations, Export Control Laws as appropriate and will have access to only unclassified data sets

- Ineligible Organizations
  - Other Government Agencies; Federally Funded Research and Development Centers (including the National Laboratories); University Affiliated Research Centers

Proposal Evaluation Criteria

Evaluation Criteria:

- Overall scientific and technical merit
- Effectiveness of work plan
- Relevance to IARPA mission
- Relevant experience and expertise
- Cost realism
- Security (Pass/Fail) – Prime must have a S//NF facility clearance and personnel cleared at the S//NF level

Meeting and Travel Requirements

Base Contract and each Option Year:

- Program kick-off meeting with each contractor team
  - Site visit at prime contractor’s facility with contractor’s entire team
- Site visits by Program Office to each prime contractor facility prior to evaluation
  - First visit: review research plans
  - Second visit: contractor provides demonstration of research prototype
- System Evaluation
  - Evaluations will take place at government-specified test facilities
  - We anticipate one West Coast test facility and one East Coast test facility
  - Several members of the contractor team will be required to participate in the evaluation
- Individual feedback sessions with each contractor team
- Workshop with all performer teams to present their results
  - Contractor teams travel to D.C. area

Publication

- **Publications will require pre-publication review**
- **Presentation in an annual KDD Program workshop will be required**
- **Regular status reports to the KDD Program Office will be required**

Questions?

Points of Contact

Dr. Arthur H. Becker
Program Manager
IARPA, Incisive Analysis Office
Office of the Director of National Intelligence
Intelligence Advanced Research Projects Activity
Washington, DC 20511

Phone: 240-373-5301
Fax: 240-373-5326
Email: dni-iarpa-baa-09-10@ugov.gov
(include IARPA-BAA-09-10 in the Subject Line)

Website: www.iarpa.gov